# Target Language-Aware Constrained Inference for Cross-lingual Dependency Parsing

Tao Meng[1], Nanyun Peng[2], Kai-Wei Chang[1]

[1]University of California, Los Angeles [2]University of Southern California

## Overview

**Task: Cross-lingual Dependency Parsing**


**Motivations**

- Prior work: focus on capturing *commonalities* between languages.
- Leverage linguistic properties of the target to facilitate the transfer.

**Contributions**

- We explore *corpus linguistic statistics* derived from WALS features and compile them into *corpus-wise constraints* to guide the inference process during the test time.
- We improve the performances on 17 out of 19 target languages.

## Background

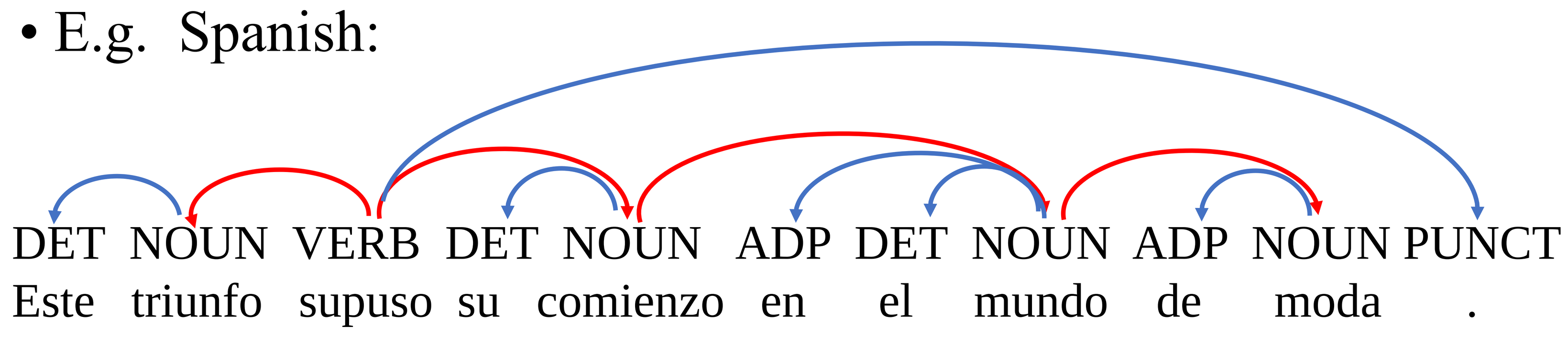
**Graph-Based Parser:**

- Assigns a score for every word pair and conducts inference to derive a directed spanning tree with the highest accumulated score.
- Integer linear program (ILP) Inference: $\max_{Y \in \mathcal{Y}} \sum_{k,i,j} S_{ij}^{(k)} y_k(i,j)$

## Corpus-Statistics Constraints

**Unary constraints:**

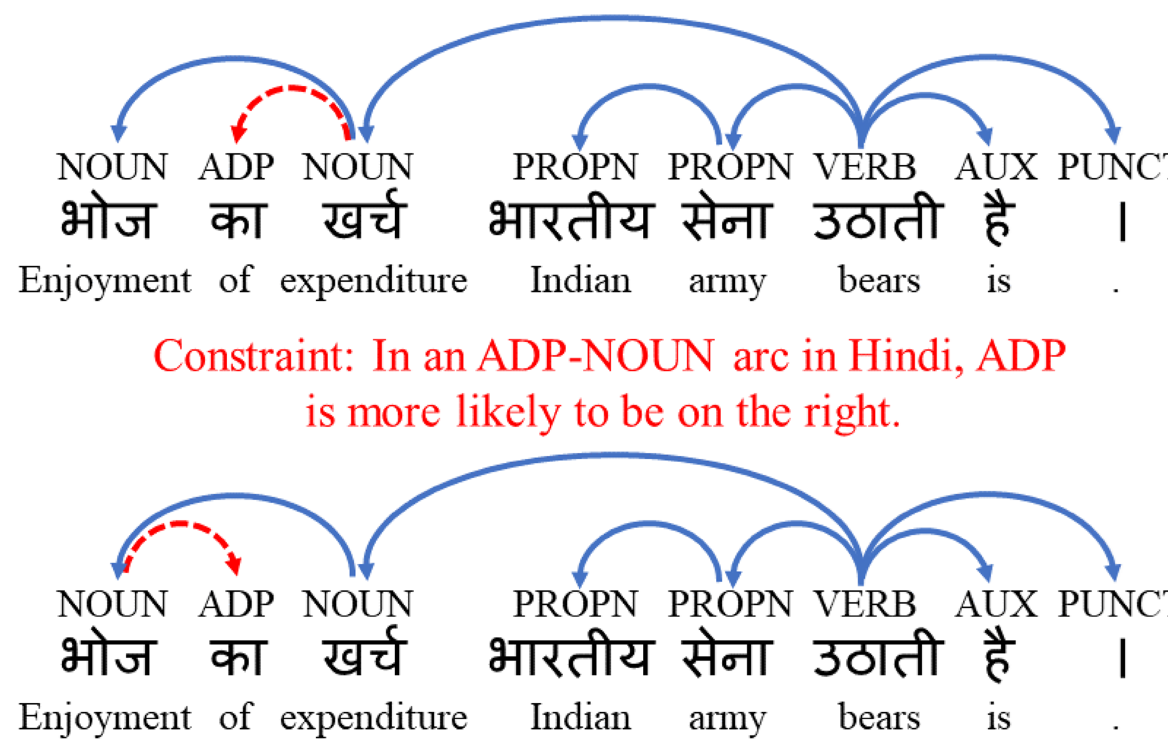
- Statistics regarding a particular POS tag ($POS$).
- E.g. Spanish:




Heads of NOUN appears on the left 82.9% of the time.

**Binary constraints:**

- Statistics regarding a pair of POS tags ($POS_1, POS_2$).
- E.g. In Hindi, ADP appears on the right of NOUN in ADP-NOUN arcs 99.9% of the time




- Given parse trees $Y$ and a constraint $C$, we define the ratio function $R(C, Y)$.

$$R(C,Y) := \frac{\sum_k \sum_{i,j:(k,i,j) \in C^+} y_k(i,j)}{\sum_k \sum_{i,j:(k,i,j) \in C^+ \cup C^-} y_k(i,j)},$$

- Constraints: statistics of Y consistent with the pre-defined ratio $r$:

$$r - \theta \le R(C, Y) \le r + \theta. \qquad \theta\text{: pre-defined margin}$$

- WALS features → three types of constraints:

$$WALS_{NOUN} \xrightarrow{LinearRegression} C1 = (NOUN),$$

$$WALS_{85A} \rightarrow C2 = (NOUN, ADP),$$

$$WALS_{87A} \rightarrow C3 = (NOUN, ADJ).$$

- Dominant order → 75% or more.

## Inference with Corpus-Statistics Constraints

- **Lagrangian Relaxation (Right).**
  - Constrained inference problem can be formulated as an ILP:
    $\max_{Y \in \mathcal{Y}} \sum_{k,i,j} S_{ij}^{(k)} y_k(i,j)$. s.t. $r_i - \theta_i \le R(C_i, Y) \le r_i + \theta_i,\ i \in [N]$
  - Solve approximately by Lagrangian Relaxation:
    - Lagrangian multipliers $\boldsymbol{\lambda}$ → relax the constraints.
    - Iteratively ($\boldsymbol{\lambda}^{(t)} \xrightarrow{Inference} Y \xrightarrow{Gradient} \boldsymbol{\lambda}^{(t+1)}$)
    - Inference with the trained multipliers $\boldsymbol{\lambda}^{(T)}$.
- **Posterior Regularization (Middle).**
  - Treat the model as a probability model $p_\theta$:

$$p_k(i,j) \propto \exp S_{ij}^{(k)}$$

  - Define the feasible set $Q$ by constraints:

$$r_i - \theta_i \le R(C_i, q) \le r_i + \theta_i, i \in [N]$$

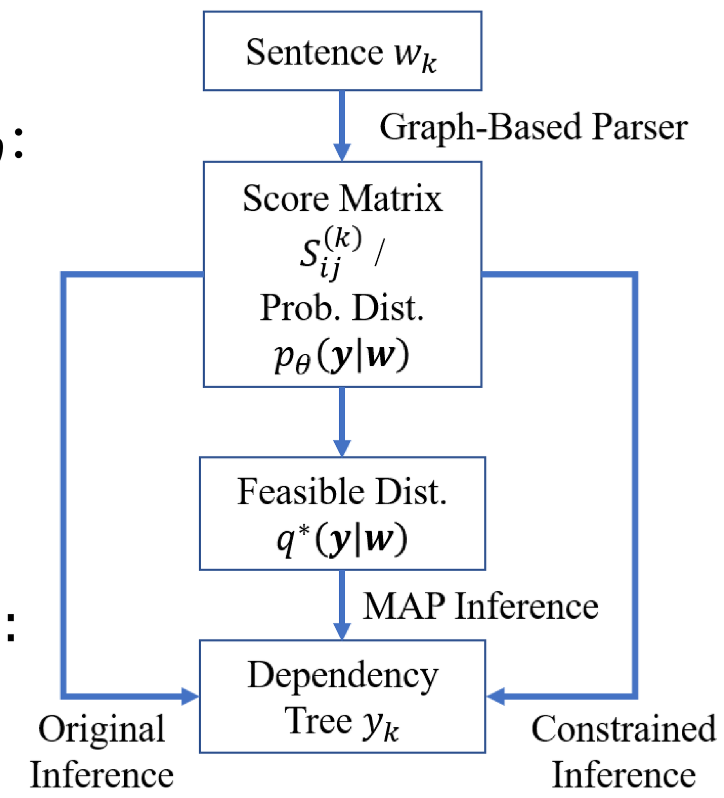
  - Find the closest distribution in $Q$ from $p_\theta$:

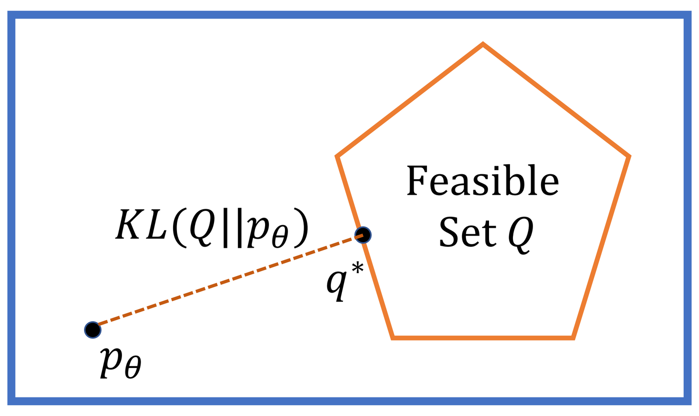
$$q^* = \arg\min_{q \in Q} KL(q || p_\theta)$$

  - MAP inference based on the feasible distribution $q^*$.

$$Y = \arg\max_{Y \in \mathcal{Y}} q^*(Y)$$

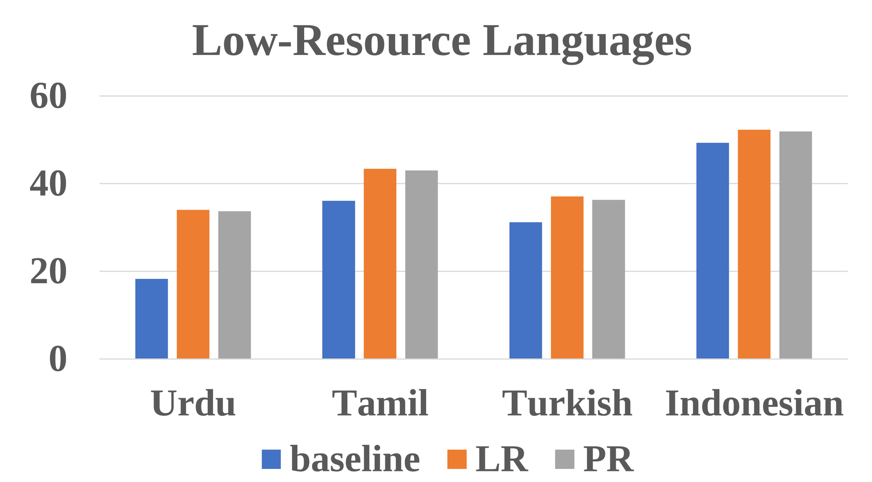
$$= \arg\max_{Y \in \mathcal{Y}} \prod_{k,i,j} q_k^*(i,j)^{y_k(i,j)}$$

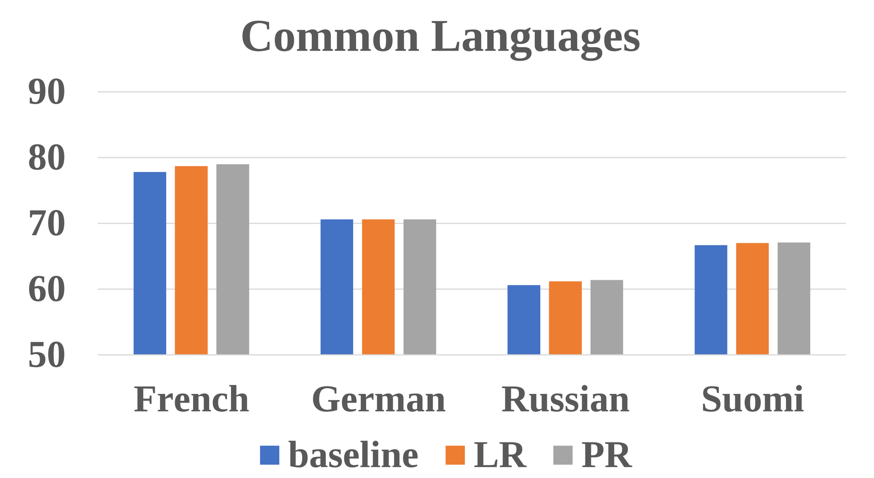



## Results

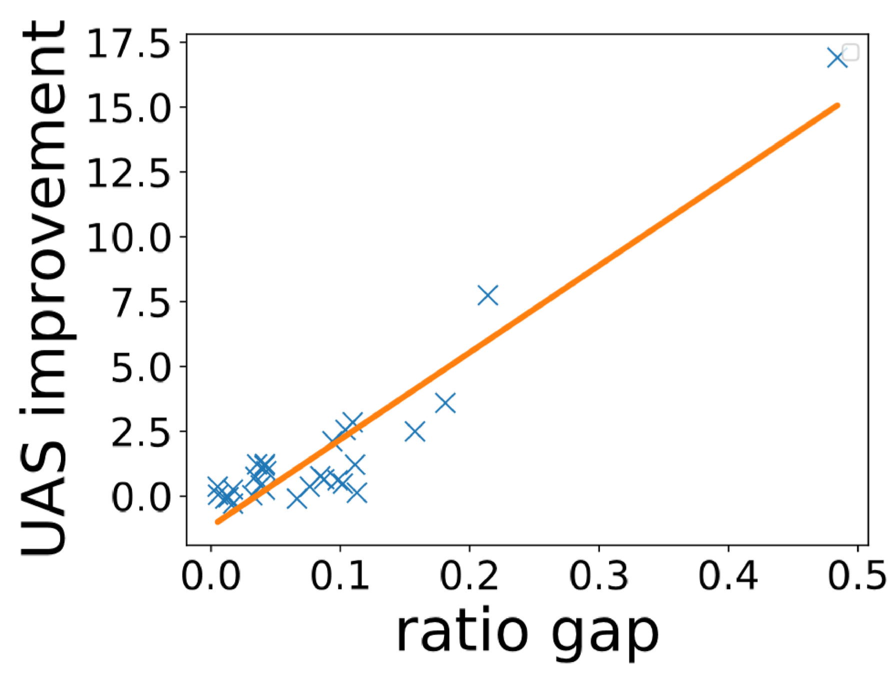
- *Significant improvements* in low-resource languages. Keep or slightly improve the performance in common languages.




- Analysis about individual constraints and the relation between improvements and ratio gap (*Highly related*, *Pearson 0.938*).

| Model | UAS | coverage | Δ |
|---|---|---|---|
| baseline | 54.3 | N/A | N/A |
| +Proj. | 54.6 | N/A | +0.3 |
| +Proj.+C1 | 57.0 | 0.24 | +2.4 |
| +Proj.+C2 | 55.7 | 0.08 | +1.1 |
| +Proj.+C3 | 55.0 | 0.07 | +0.4 |
| oracle | 58.4 | N/A | +4.1 |




## Conclusion

- Improve *15* and *17* languages out of 19 with LR and PR, respectively.
- Languages with different word order from English improve significantly.
- Lagrangian relaxation has a greater average improvement, while posterior regularization improves more languages.
- Code and models:
  https://github.com/MtSomeThree/CrossLingualDependencyParsing/